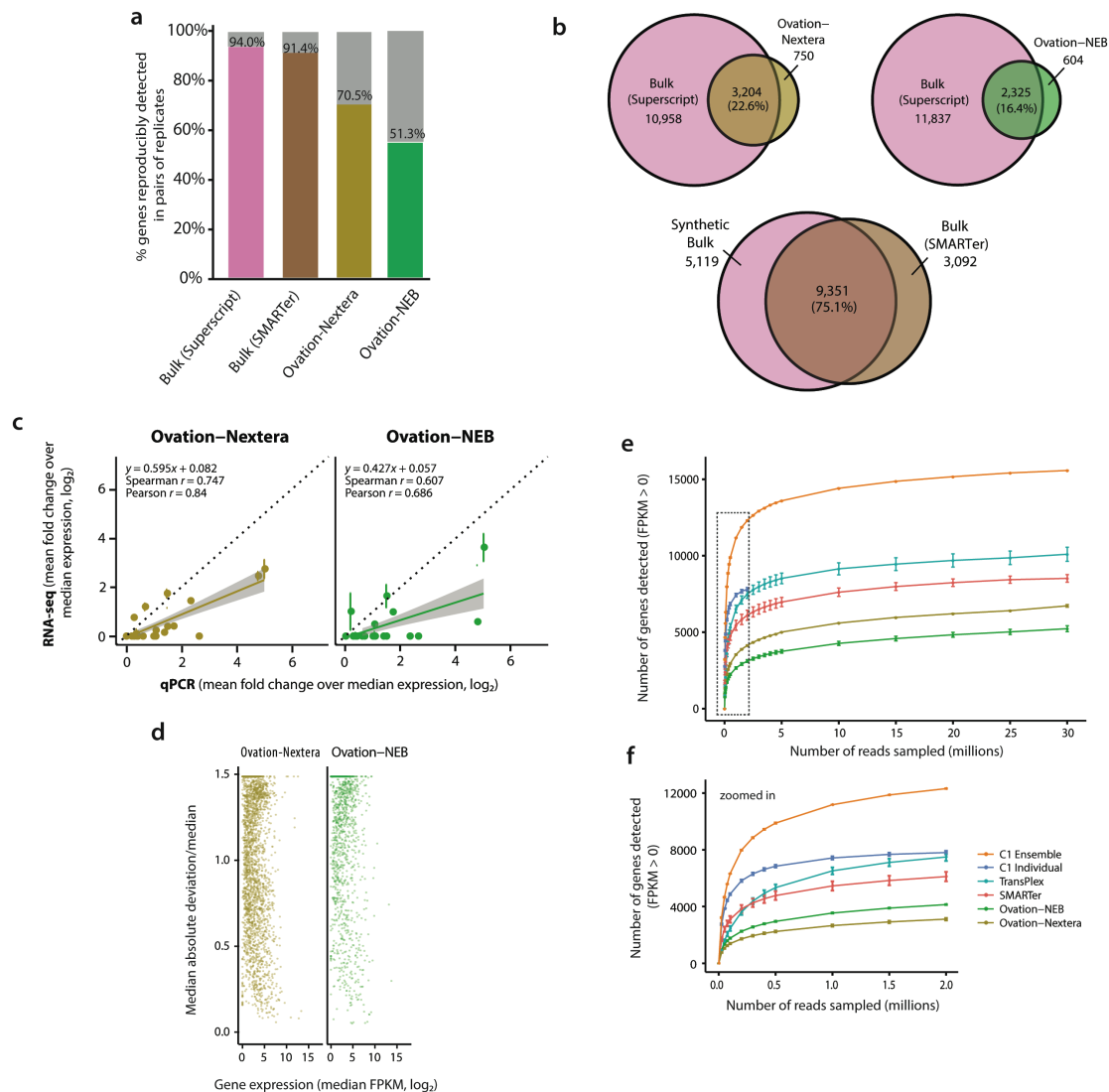


High throughput quantitative whole transcriptome analysis from single cells

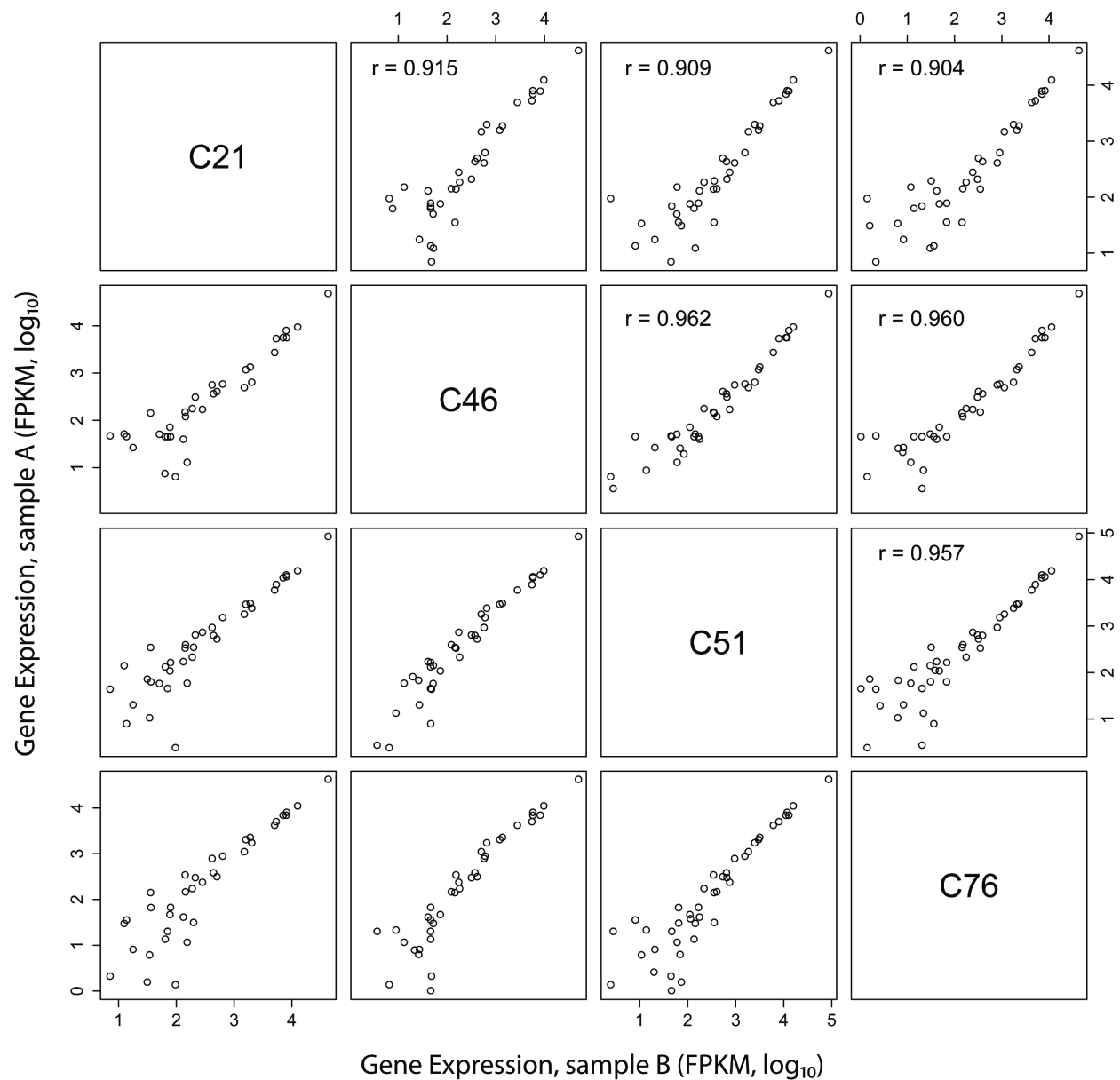
Angela R. Wu, Norma F. Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E. Rothenberg, Francis M. Mburu, Gary L. Mantalas, Sopheak Sim, Michael F. Clarke, Stephen R. Quake

Supplementary Figure 1 | Results using Ovation for cDNA synthesis and amplification.



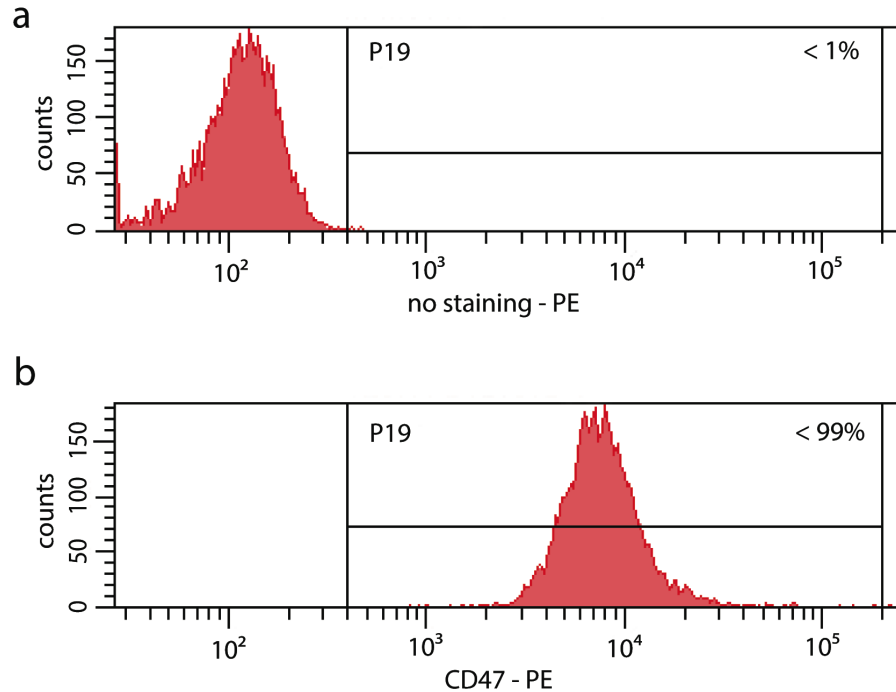
This data was generated using Nugen Ovation version 1 kit. Since the research was performed, an updated version 2 kit has been released. **(a)** Assessment of reproducibility (same as Fig. 1b). **(b)** Assessment of sensitivity relative to bulk (same as Fig. 1c). **(c)** Correlation between single-cell RNA-seq and single-cell multiplexed qPCR for Ovation methods (same as Fig. 2). **(d)** Variation in the measured gene expression as a function of gene expression level across sample replicates (same as Fig. 5b). **(e)** Saturation curves for Ovation methods (same as Fig. 5a). **(f)** Saturation curves, zoomed in 0-2 million reads, for Ovation methods (same as Fig. 5b).

Supplementary Figure 2 | Assessment of reproducibility within microfluidically generated datasets.



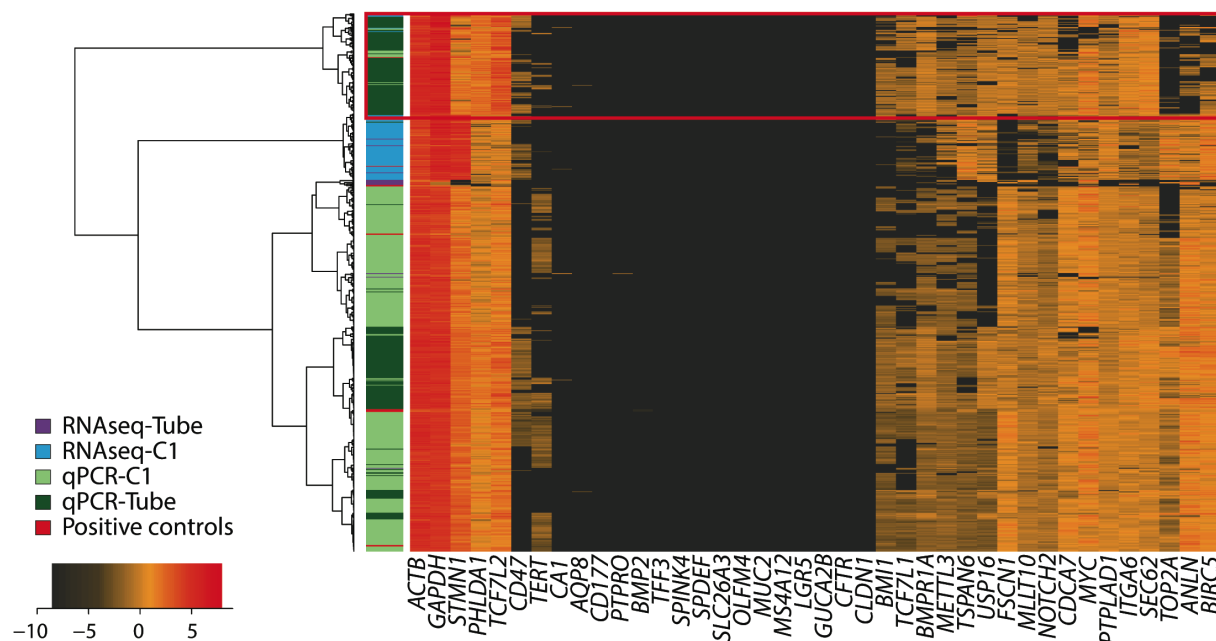
Correlation between the transcript levels of all ERCC spike-ins between pairs of randomly selected single cell samples. Transcript level is measured in FPKM, and is plotted on a log₁₀ transformed scale. The Pearson correlation coefficient for each pair is calculated and noted in each plot. The high degree of correlation between sample pairs indicates that transcript quantities are consistently detected in multiple samples, demonstrating reproducibility of this method in measuring gene expression levels.

Supplementary Figure 3 | Analysis of CD47 protein expression in HCT116 cells.



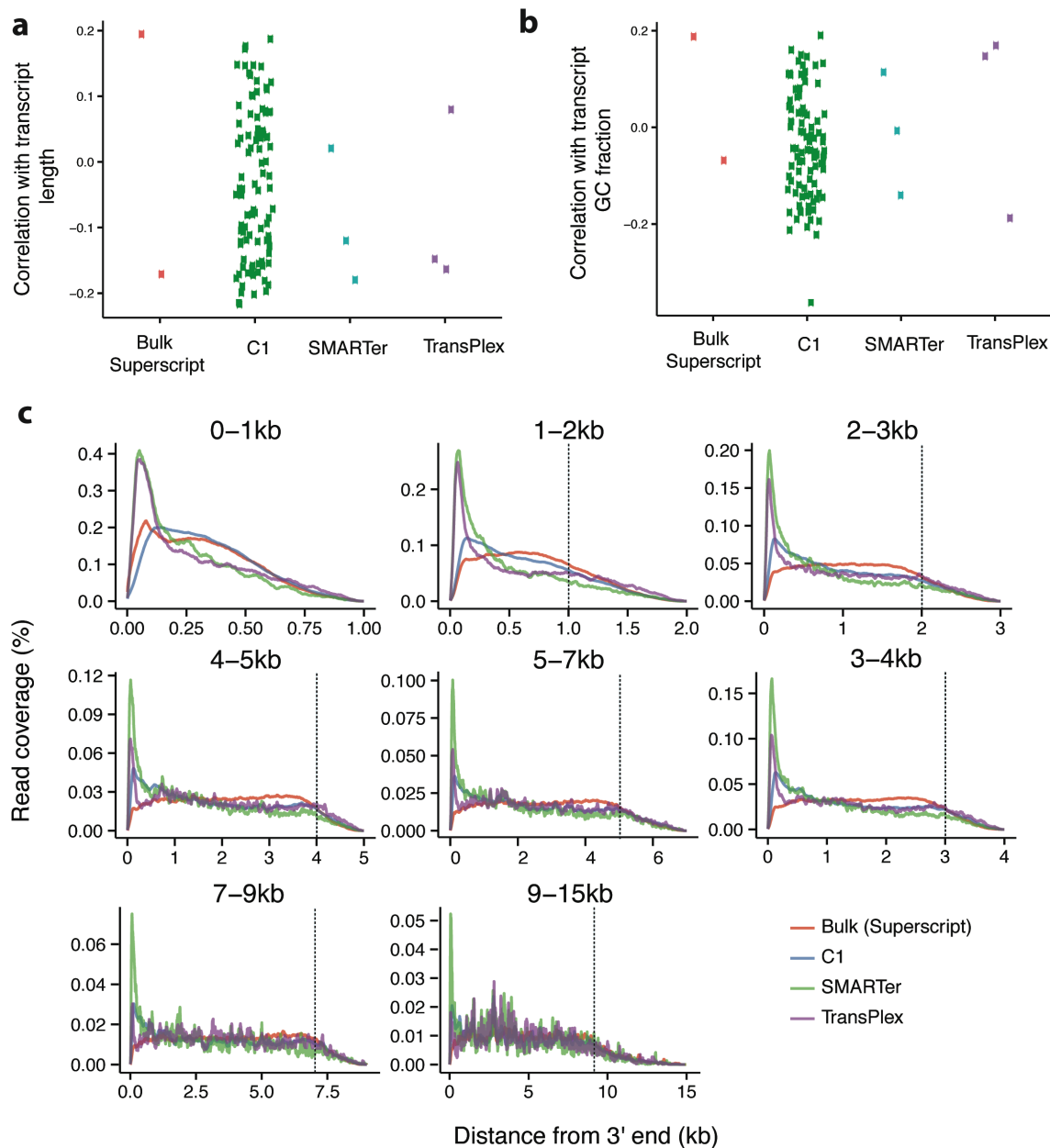
Surface expression of the CD47 protein was evaluated in HCT116 cells by flow cytometry, using a mouse anti-human CD47 monoclonal antibody conjugated to phycoerythrin (PE). **(a)** Analysis of the baseline fluorescence of unstained HCT116 cells and definition of the threshold used to separate negative from positive events in the PE fluorescence channel. The threshold is set by creation of an analysis gate (P19), which encompasses < 1% of unstained cells with the highest baseline fluorescence levels. **(b)** Analysis of HCT116 cell fluorescence after staining with the anti-CD47 monoclonal antibody (clone B6H12; BD Biosciences). The percentage of CD47⁺ cells is calculated as the percentage of cells contained within the P19 gate. The experiment reveals high levels of CD47 protein expression across the whole HCT116 population (>99% of analyzed cells).

Supplementary Figure 4 | Dendrogram representing unsupervised hierarchical clustering of gene expression of 40 genes for HCT116 cells.



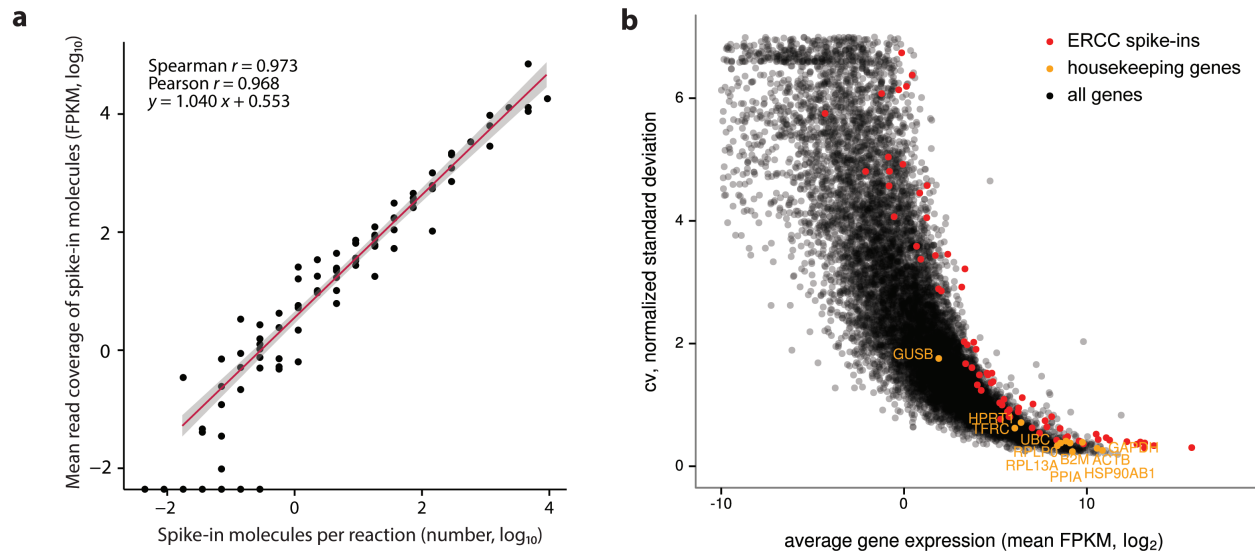
Clustering was done across all samples, color-coded by sample reaction volume (prepared in microliter or nanoliter reaction volumes) and measurement method (RNA-seq or qPCR). Fold-changes in expression over median gene expression for each sample were used for clustering. Genes known to be unexpressed in HCT116 are found to have no expression or very low expression in all experiments, and highly expressed genes were consistently represented with both methods. Cells whose gene expressions were derived using the C1 system are interspersed with those cells measured using microliter volume qPCR, indicating that the two methods are similar. The bulk RNA positive controls are also not clustered together, indicating that in general, gene expression patterns are similar between the single-cell and bulk samples for these selected genes. Some of the cells used in this experiment were harvested during semi-confluency where more cells are proliferating, and others were harvested at confluency with cells being in a slower growth state. Red box highlights a subset of cells characterized by lower expression levels of the ANLN, TOP2A and BIRC5 genes, likely representing more quiescent, non-dividing cells²⁷⁻²⁹ (unpublished data, TK, PD, SS, MFC, SRQ).

Supplementary Figure 5 | Analysis of coverage bias for each sample preparation method.



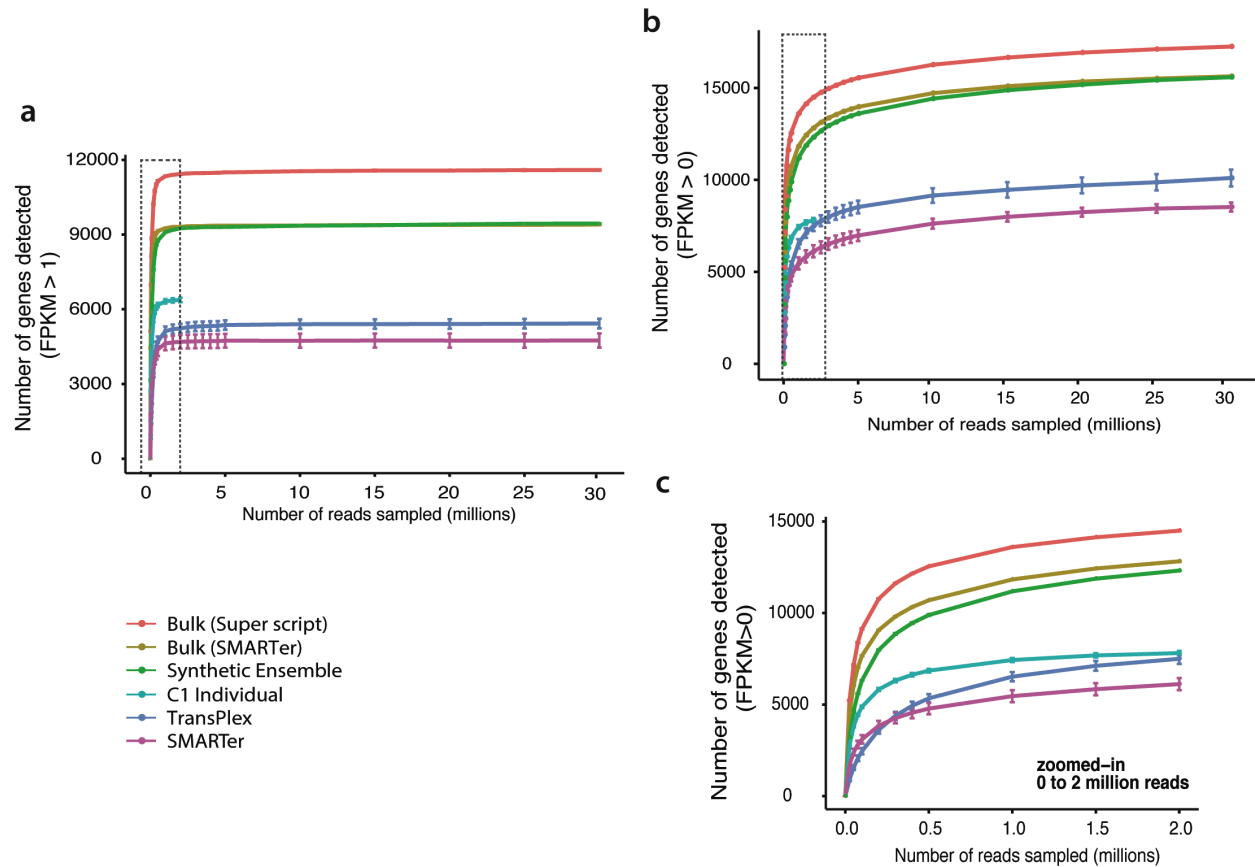
(a) Pearson correlation coefficients between \log_{10} transformed expression (FPKM) and transcript length, compared between all methods. Each color represents a replicate using that method. **(b)** Pearson correlation coefficients between \log_{10} transformed expression (FPKM) and transcript GC content, compared between all methods. No systematic biases are observed between expression level and transcript length or GC content. **(c)** Comparison of read coverage over the length of the transcript between different sample preparation methods. Transcripts have been segmented by length, and in each panel the read coverage is reported by distance from the 3' end. Read coverage here is defined as the number of reads covering each base divided by the total coverage over the entire transcript (i.e. percentage of total bases mapped to this transcript that covers each base). The length of the shortest included transcript is indicated by a dotted vertical line, after which a decline in read coverage is expected.

Supplementary Figure 6 | Assessment of technical variation within microfluidically generated datasets.



(a) Quantitative assessment of amplification bias and limit of detection in nanoliter volume sample preparation for single-cell RNA-seq. The concentration of each exogenously spiked-in transcript is known, and plotted here against their corresponding mean read coverage as measured by RNA-seq. The concentrations are \log_{10} -transformed molecules per volume. Read coverage is represented by the \log_{10} -transformed mean FPKM value across all single-cell samples for each transcript. Each data point represents the quantitation of a particular transcript in the exogenous transcript mixture and is an average value over all 96 chambers, each of which represents an independent replicate. **(b)** Relationship between mean expression level across all single cells and the coefficient of variation for all genes in HCT116 cells as well as for exogenous ERCC RNA spike-ins, which allow for empirical determination of the limit of detection and technical noise^{25,26}. Spike-in transcripts are shown in red, commonly used human housekeeping genes (Qiagen housekeeping genes PCR array) are shown in yellow, and other endogenous genes are shown in black. Genes with high biological variation are expected to show higher variability than expected from pure technical noise for a given average gene expression as described by the ERCC spike-ins. The fact that only two genes exhibit higher variation than the ERCC spike-ins confirms the homogeneity of the HCT116 cell line.

Supplementary Figure 7 | Saturation curves for various methods (FPKM > 0).



(a) Saturation curve from 0 to 30 million reads, by sample preparation method. Each point on the curve was generated by randomly selecting the corresponding number of millions of raw reads from each sample library, and then using the same alignment pipeline to call genes. This random sub-sampling was repeated for each sample replicate for a total of 4 sub-sampled datasets per point, and the mean number of genes with FPKM greater than 1 is plotted. The C1 samples individually were only sequenced to a depth of 2 million reads on average; therefore no data points beyond 2 million reads were created for those samples. **(b)** Saturation curve from 0 to 30 million reads, by sample preparation method. This plot is similar to panel (a) of this figure, but the number of genes detected at each sequencing depth is expanded to include low abundance genes that have FPKM between 0 and 1. Error bars show standard error over the four sub-sampled datasets. **(c)** Saturation curve from 0 to 2 million reads, by sample preparation method, for genes with FPKM values greater than 0. Similar to Figure 5.

Supplementary Table 1 | Summary of RNA-seq experiments and their basic cDNA synthesis mechanism.

Method	cDNA synthesis	Library construction	# Samples
Bulk RNA	1) Magnetic bead-based oligo dT priming extraction of mRNA from cell lysates, followed by Superscript II cDNA synthesis	Nextera – tagmentation using transposase enzymes	1) $n = 2$
	2) SMARTer Ultra Low RNA kit – oligo dT priming		2) $n = 2$
Clontech SMARTer	SMARTer Ultra Low RNA Kit – oligo dT priming	Nextera	$n = 3$
Sigma TransPlex	Sigma-Aldrich TransPlex WTA kit – random priming for both first and second strand synthesis, with a universal 5' priming sequence for subsequent PCR amplification.	Nextera	$n = 3$
NuGEN Ovation - Nextera	NuGEN Ovation RNA-seq kit (v1)	Nextera	$n = 3$
NuGEN Ovation - NEBNext	NuGEN Ovation RNA-seq kit (v1)	NEBNext Library Prep kit	$n = 4$

Supplementary Table 2 | List of primers used for qPCR.

GENE	Gene chr span	refseq	Assay length	ABI Taqman Assay ID
ACTB--333	Chr.7: 5566779 - 5570232	NM_001101.3	63	Hs00357333_g1
ANLN--612	Chr.7: 36429432 - 36493400	NM_018685.2	71	Hs01122612_m1
AQP8--279	Chr.16: 25228285 - 25240253	NM_001169.2	57	Hs01086279_m1
BIRC5--353	Chr.17: 76210277 - 76221716	NM_001012271.1	93	Hs00153353_m1
BMI1--411	Chr.10: 22610006 - 22620414	NM_005180.8	105	Hs00180411_m1
BMP2--564	Chr.20: 6748745 - 6760911	NM_001200.2	84	Hs01055564_m1
BMPR1A--913	Chr.10: 88516396 - 88684945	NM_004329.2	94	Hs01034913_g1
CA1--139	Chr.8: 86240458 - 86290342	NM_001128829.2	85	Hs00266139_m1
CD177--669	Chr.19: 43857825 - 43867480	NM_020406.2	66	Hs00360669_m1
CD47--37	Chr.3: 107761941 - 107809935	NM_001777.3	131	Hs00963737_m1
CDCA7--242	Chr.2: 174219561 - 174233718	NM_031942.4	81	Hs00912242_g1
CFTR--537	Chr.7: 117120017 - 117308719	NM_000492.3	69	Hs01565537_m1
CLDN1--357	Chr.3: 190023490 - 190040235	NM_021101.4	71	Hs01076357_m1
FSCN1--051	Chr.7: 5632454 - 5646286	NM_003088.3	126	Hs00602051_mH
GAPDH--905	Chr.12: 6643657 - 6647536	NM_002046.4	122	Hs99999905_m1
GUCA2B--189	Chr.1: 42619092 - 42621495	NM_007102.2	59	Hs00951189_m1
ITGA6--011	Chr.2: 173292314 - 173371181	NM_000210.2	64	Hs01041011_m1
LGR5--421	Chr.12: 71833813 - 71978622	NM_003667.2	78	Hs00969421_m1
METTL3--158	Chr.14: 21966282 - 21979457	NM_019852.3	87	Hs01096158_m1
MLLT10--021	Chr.10: 21823102 - 22032554	NM_001195626.1	66	Hs00946021_m1
MS4A12--572	Chr.11: 60260251 - 60274903	NM_001164470.1	80	Hs00214572_m1
MUC2--094	Chr.11: 1074875 - 1104417	NM_002457.2	64	Hs03005094_m1
MYC--408	Chr.8: 128748315 - 128753680	NM_002467.4	107	Hs00153408_m1
NOTCH2--747	Chr.1: 120454176 - 120612276	NM_024408.3	73	Hs00225747_m1
OLFM4--437	Chr.13: 53602972 - 53626192	NM_006418.4	85	Hs00197437_m1
PHLDA1--810	Chr.12: 76419227 - 76425556	NM_007350.3	78	Hs00705810_s1
PTPLAD1--905	Chr.15: 65822827 - 65870693	NM_016395.2	60	Hs01012905_m1
PTPRO--097	Chr.12: 15475487 - 15750335	NM_002848.3	102	Hs00243097_m1
SEC62--753	Chr.3: 169684580 - 169716161	NM_003262.3	87	Hs00963753_m1
SLC26A3--365	Chr.7: 107405912 - 107443678	NM_000111.2	66	Hs00995365_m1
SPDEF--942	Chr.6: 34505580 - 34524091	NM_001252294.1	68	Hs00171942_m1
SPINK4--780	Chr.9: 33240196 - 33248565	NM_014471.1	66	Hs01018780_m1
STMN1--370	Chr.1: 26210677 - 26233368	NM_001145454.1	104	Hs00606370_m1
TCF7L1--103	Chr.2: 85360734 - 85537505	NM_031283.2	65	Hs01064103_m1
TCF7L2--053	Chr.10: 114710009 - 114927437	NM_001146274.1	76	Hs01009053_m1
TERT--656	Chr.5: 1253282 - 1295162	NM_001193376.1	79	Hs00972656_m1
TFF3--625	Chr.21: 43731777 - 43735706	NM_003226.3	98	Hs00173625_m1
TOP2A--137	Chr.17: 38544773 - 38574202	NM_001067.3	81	Hs01032137_m1
TSPAN6--458	Chr.X: 99883795 - 99891794	NM_003270.2	95	Hs01073458_m1
USP16--191	Chr.21: 30396938 - 30426809	NM_001001992.1	116	Hs01062191_m1

Supplementary Note | Calculation of detection rate for C1 microfluidic single-cell RNA-seq

To estimate the probability of detecting the transcript if the transcript concentration is at 1 molecule per chamber, we first determine this concentration in attomoles/nL, based on the microfluidic device geometry:

Volume of the lysis chamber in the device is ~9 nL
1 molecule per chamber = 1 molecule per 9 nL
= 0.11 molecule/nL
= 110 molecules/ μ L

The ERCC spike-ins were diluted 40,000X from stock when added to reaction, so for a transcript to be at 1 molecule per chamber in the device, the original stock concentration would have to be:

$$\begin{aligned} & 110 \times 40,000 \text{ molecules}/\mu\text{L} \\ & = 4,400,000 \text{ molecules}/\mu\text{L} \\ & = 7.30637 \text{ attomoles}/\mu\text{L} \end{aligned}$$

In mix A, there are five ERCC transcripts that are at a concentration of 7.324 attomole/ μ L: ERCC-00034, ERCC-00085, ERCC-00154, ERCC-00157, and ERCC-00160

An R script was then used to extract measured FPKM abundances for these five transcripts from each single cell sample (i.e. each chamber), and the number of non-zero measurements was counted. This number divided by the total number of chambers (96) gives the probability of detection for each transcript at this concentration. The reported value of 0.4 in the manuscript is the arithmetic mean taken across all five of such transcripts.